Universidade de Brasília
Instituto de Ciências Exatas
Departamento de Estatística

IMPLEMENTAÇÃO DE ALGORITMO COMPUTACIONAL PARA ANÁLISE DE DADOS AMOSTRAIS COMPLEXOS

João Renato Falcão **10/0107575**

Brasília

2013

João Renato Falcão **10/0107575**

IMPLEMENTAÇÃO DE ALGORITMO COMPUTACIONAL PARA ANÁLISE DE DADOS AMOSTRAIS COMPLEXOS

Relatório apresentado à disciplina Estágio Supervisionado II do curso de graduação em Estatística, Departamento de Estatística, Instituto de Exatas, Universidade de Brasília, como parte dos requisitos necessários para o grau de Bacharel em Estatística.

Orientador: Prof. Dr. Alan Ricardo da Silva

Brasília

2013

Agradecimentos

Ao SAS *Institute* Brasil por possibilitar a utilização desse software por meio de parceria acadêmica com o Departamento de Estatística da UnB. Ao Instituto de Pesquisa Econômica Aplicada - IPEA pelo banco de dados da PNAD com os dicionários de variáveis.

Agradeço à minha família pelo apoio, atenção e paciência destinados a mim em todos os momentos cruciais deste período. Aos colegas de curso cuja companhia e amizade me ajudaram em muitos momentos de decisão. Aos professores por todo conhecimento transmitido e em especial ao meu professor orientador, Alan Ricardo da Silva, pela ajuda, extrema paciência diante das dificuldades que eu tive e tempo dedicados a este trabalho. Um agradecimento especial à Laíssa, por todo o incentivo, amor e carinho, os quais me impulsionaram a tentar obter os melhores resultados.

João Renato Falcão

Resumo

Toda análise de dados obtidos via amostragem complexa deve ser feita considerando o método de seleção, pois estimativas pontuais de parâmetros da população são influenciadas por pesos distintos das observações, já estimativas de variância são influenciadas por conglomeração, estratificação e pesos. Com isso, o trabalho procurou verificar o viés causado pela não incorporação do plano amostral na análise, a influência que este exerce no erro padrão dos estimadores de parâmetros de regressão e elaborar uma algoritmo capaz de incluir os três estágios de seleção da PNAD em apenas uma *macro* do SAS. A partir dos resultados obtidos pôde-se observar que cada estágio atua na variância total, porém o terceiro pode ser omitido por ser quase imperceptível no resultado. O algoritmo mostrou-se eficaz, obtendo com precisão os coeficientes estimados e seus erros padrão.

Lista de Tabelas

Lista de Figuras

Sumário

Capítulo 1

INTRODUÇÃO

Toda informação obtida por métodos de amostragem complexa deve ser analisada considerando o método de coleta. Isso se deve ao fato de que as estimativas pontuais de parâmetros da população são influenciadas por pesos distintos das observações e as estimativas de variância são influenciadas pela conglomeração, estratificação e pesos (Pessoa and Silva, 1998).

Contudo, os pacotes estatísticos usualmente utilizados para se fazer estudos analíticos baseados em amostra consideram as observações como sendo independentes e identicamente distribuídas. Isso significa dizer que a análise feita utilizando um software usual irá considerar que o plano amostral adotado foi o de amostragem aleatória simples.

Ao ignorar aspectos como probabilidades distintas de seleção das unidades, conglomeração, estratificação, entre outros, os pacotes estatísticos tradicionais podem produzir estimativas incorretas das variâncias das estimativas pontuais (Pessoa and Silva, 1998).

Para avaliar o impacto do plano amostral na inferência utiliza-se o conceito de Efeito do Plano Amostral (EPA), o qual consiste na razão entre variâncias calculadas

considerando o plano amostral adotado e considerando que os dados foram obtidos por amostragem aleatória simples.

No *software* SAS, a maior parte dos procedimentos usuais tais como MEANS, FREQ, GLM, etc, são capazes de calcular médias amostrais e/ou estimar parâmetros de regressão, porém partem do pressuposto de que a população analisada é infinita e a amostra foi retirada por amostragem aleatória simples (SAS, 2012). Portanto, ao se analisar dados cuja amostra foi obtida com método de amostragem complexa, como é o caso da PNAD (Pesquisa Nacional por Amostra de Domicilio), tais procedimentos podem não ser adequados. O resultado disso é um possível viés na inferência e uma super ou subestimação da variância para o caso de dados amostrais complexos.

A partir da versão 8.2, o SAS apresenta como alternativa procedimentos que consideram na análise o plano amostral adotado. São eles: SURVEYMEANS, SURVEYFREQ, SURVEYREG, SURVEYLOGISTIC e SURVEYPHREG, podendo ser utilizados tanto para casos de um ou mais estágios de seleção, com ou sem reposição e pesos desiguais (SAS, 2012). Entretanto, não há como fazer uma análise de dados com mais de um estágio de seleção em um procedimento apenas, sendo necessário executar um procedimento SURVEY para cada estágio.

Dessa forma o objetivo é implementar um algoritmo único para análise de dados com três estágios de seleção, como é o caso da PNAD. Assim, o usuário não precisará conhecer o desenho amostral utilizado para fazer a análise dos dados.

1.1 OBJETIVOS

O objetivo geral do trabalho é implementar um algoritmo no *software* SAS que analise dados complexos (amostragem em 3 estágios) em apenas um procedimento. Os objetivos específicos são:

- Identificação das características de um plano amostral complexo;
- Desenvolvimento de algoritmo computacional no SAS/IML para estimação dos parâmetros da regressão linear, incorporando o plano amostral complexo;
- Avaliar o viés provocado nos resultados da análise de regressão sem considerar o plano amostral.

Capítulo 2

PLANOS AMOSTRAIS

2.1 INTRODUÇÃO

Os planos amostrais consistem de diferentes técnicas para levantamento de dados de forma a selecionar uma parcela de elementos grande o suficiente apenas para ser representativa da população e minimizar possíveis vieses no resultado da análise. Cada plano amostral é uma forma diferente de coleta onde se conhece previamente a probabilidade de se coletar cada indivíduo dentro da população, fazendo com que seja possível se estimar os parâmetros e suas respectivas variâncias.

Existem três tipos básicos de amostragem: Amostragem Aleatória Simples, Amostragem por Conglomerado e Amostragem Estratificada. Cada técnica é usada de acordo com a maneira como está disposta a população a ser estudada, de modo que dependendo da situação uma técnica pode ou não ser mais vantajosa que a outra. Suas definições e respectivas situações de uso são definidas a seguir.

2.2 Amostragem Aleatória Simples

Utiliza-se o método de Amostragem Aleatória Simples (AAS) quando os dados estão dispersos de maneira homogênea, ou seja, qualquer amostra coletada será re-

presentativa do universo onde cada elemento tem a mesma chance de ser selecionado. Sendo assim, essa técnica consiste em sortear com ou sem reposição uma parcela representativa de indivíduos da população. O método abordado aqui será sempre sem reposição, no qual apesar de cada indivíduo ter a mesma chance de ser selecionado antes do sorteio, essa probabilidade muda para cada elemento escolhido.

A média amostral é dada por:

$$\bar{x} = \sum_{i=1}^{n} \frac{x_i}{n} \tag{2.1}$$

A média amostral é um estimador não viesado para a média populacional (Cochran, 1977).

A variância para uma amostra aleatória simples com reposição (AAS_c) é dada por (Cochran, 1977):

$$\sigma^2 = \frac{\sum_{i=1}^{N}(x_i - \bar{X})^2}{N} \tag{2.2}$$

Ainda segundo Cochran (1977), a variância para uma amostra aleatória simples sem reposição (AAS_s) é dada por:

$$S^2 = \frac{\sum_{i=1}^{N}(x_i - \bar{X})^2}{N-1} \tag{2.3}$$

Sendo assim, definem-se as variâncias da média $\bar{x}$ como (Cochran, 1977):

$$Var_{AAS_c}(\bar{x}) = \frac{\sigma^2}{n} \tag{2.4}$$

$$Var_{AAS_s}(\bar{x}) = \frac{S^2}{n}\frac{(N-n)}{N} = \frac{S^2}{n}(1-f) \tag{2.5}$$

onde f é dado por $\frac{n}{N}$.

Essa proporção inserida na fórmula é conhecida como fator de Correção para População Finita (CPF), ou do inglês *Finite Population Correction* (FPC) (Cochran, 1977).

É válido observar que para o cálculo dessa variância é preciso conhecer previamente alguns parâmetros populacionais tais como seu tamanho e a média de seus valores. Na prática, tais parâmetros não podem ser conhecidos, mas podem ser estimados a partir dos dados amostrais (Cochran, 1977).

De acordo com Cochran (1977), um estimador não viesado da variância populacional estimada S^2 ou σ^2 é dado por:

$$\widehat{Var}(\bar{x}) = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1} = s^2 \tag{2.6}$$

2.3 Amostragem Aleatória Estratificada

Quando a seleção de uma amostra por meio de AAS implica em um custo muito alto e um dispêndio de tempo maior, é necessário partir para outras técnicas de seleção. Caso a população possa ser subdividida em grupos distintos entre si, mas cujos elementos dentro de cada grupo são homogêneos, faz-se uso da técnica conhecida por Amostragem Estratificada (AE), onde cada grupo é definido como um estrato. De cada estrato, seleciona-se por AAS uma amostra de indivíduos e assim, todas as amostras dos estratos juntas formam uma amostra representativa da população.

Para o cálculo de um estimador não-viesado da média dos valores de uma população estratificada, utiliza-se dois métodos. O primeiro consiste em calcular diretamente o estimador da média por meio de parâmetros populacionais e o segundo

com base na média da amostra. Ambos os métodos são mostrados nas equações a seguir (Cochran, 1977):

$$\bar{x}_{st} = \sum_{h=1}^{L} \frac{N_h \bar{x}_h}{N} = \sum_{h=1}^{L} W_h \bar{x}_h \tag{2.7}$$

$$\bar{x}_h = \sum_{i=1}^{n_h} \frac{x_{hi}}{n_h} \tag{2.8}$$

Considerando uma população dividida em L estratos, de tal forma que cada estrato possui N_h elementos, a Equação (2.7) utiliza parâmetros populacionais onde N é o total de elementos na população, N_h é o total de elementos no h-ésimo estrato e $W_h = \frac{N_h}{N}$ é definido como o peso do estrato na população (Cochran, 1977).

Já na Equação (2.8), calcula-se a média obtida na amostra provinda de uma população estratificada, onde n_h corresponde ao total de elementos amostrais no h-ésimo estrato, n o total de elementos na amostra e $\bar{x}_h$ a média dos valores encontrados no h-ésimo estrato.

Pode-se notar que $\bar{x}_{st}$ coincidirá com $\bar{x}$ quando os pesos na amostra forem iguais aos pesos na população, ou seja, quando a fração da amostra for a mesma em todos os estratos, o que também é conhecido como alocação proporcional de n_h (Cochran, 1977).

O cálculo da variância do estimador $\bar{x}_{st}$ é dado por (Cochran, 1977):

$$Var(\bar{x}_{st}) = \sum_{h=1}^{L} W_h^2 Var(\bar{x}_h) \tag{2.9}$$

onde $Var(\bar{x}_h)$ é a variância do estimador $\bar{x}_h$.

Uma forma mais completa de se definir a variância do estimador $\bar{x}_{st}$ é de acordo

com:

$$Var(\bar{x}_{st}) = \frac{1}{N^2}\sum_{h=1}^{L} N_h(N_h - n_h)\frac{{S_h}^2}{n_h} = \sum_{h=1}^{L} {W_h}^2\frac{{S_h}^2}{n_h}(1 - f_h) \tag{2.10}$$

onde f_h é o fator de correção para o h-ésimo estrato, ou seja,

$$f_h = \frac{n_h}{N_h}$$

Para cada estrato, existe uma variância estimada definida como ${S_h}^2$ de acordo com a Equação (2.3). Nesse caso, seja s_h^2 o estimador de S_h^2, o estimador da variância de $\bar{x}_{st}$ é por (Cochran, 1977):

$$\widehat{Var}(\bar{x}_{st}) = \frac{1}{N^2}\sum_{h=1}^{L} N_h(N_h - n_h)\frac{s_h^2}{n_h} \tag{2.11}$$

onde h representa o h-ésimo estrato.

2.4 Amostragem por Conglomerado

Para uma população subdividida em M grupos ou unidades muito parecidos entre si, mas cujos elementos são heterogêneos o suficiente para serem considerados microrepresentações da população, define-se tais unidades como Conglomerados ou "*clusters*", em inglês, (Cochran, 1977). Nesse tipo de população, o custo para levantamento dos dados por meio de AAS é muito elevado e a movimentação para identificar os elementos no campo pode consumir muito tempo.

Considerando que selecionar grupos de unidades elementares facilita a tarefa amostral, a técnica mais adequada a ser utilizada consiste em selecionar aleatoriamente, por meio de AAS, m unidades e considerar na análise todos os indivíduos de cada grupo escolhido. Esse método é conhecido como Amostragem por Conglomerado.

Essa técnica pode ser utilizada de duas formas diferentes, dependendo de como está subdividida a população. A forma mais simples é quando a população está dividida em M conglomerados e já se pode extrair diretamente uma amostra de cada um. Diz-se nesse caso que a amostragem deve ser feita em um estágio. A segunda forma é usada quando cada conglomerado da população está subdividido em outros conglomerados. Diz-se então que a amostragem deve ser realizada em dois ou mais estágios, onde cada estágio representa uma subdivisão na população.

2.4.1 Conglomerado em um estágio - tamanhos desiguais

Cochran (1977) define essa técnica para o caso onde o número de elementos contidos em cada conglomerado difere de um grupo para outro. Para se estimar o total populacional X, primeiramente o autor define,

$$x_i = \sum_{j=1}^{M_i} x_{ij} \tag{2.12}$$

o qual corresponde a $M_i\bar{x}_i$, como sendo o valor total do i-ésimo conglomerado, onde M_i corresponde ao total de elementos populacionais no mesmo conglomerado.

A variância desse estimador é calculada de acordo com a equação definida em Cochran (1977),

$$Var(\hat{X}) = \frac{M^2(1-f)}{m} \frac{\sum_{i=1}^{M}(x_i - \bar{X})}{M-1} \tag{2.13}$$

onde

$$\hat{X} = \frac{M}{m} \sum_{i=1}^{m} x_i$$

é um estimador não-viesado do total X,

$$\bar{X} = \frac{X}{M}$$

é a média populacional por conglomerado e $f = \frac{m}{M}$.

2.4.2 Amostragem por conglomerado - em três estágios

Para o caso de uma população subdividida em conglomerados, onde cada unidade também está subdividida em outros conglomerados, tem-se que o processo de amostragem deve ser feito em dois ou mais estágios. Por exemplo, pode-se estar interessado em obter informações sobre estudantes do ensino médio da rede pública de uma determinada cidade. Cada escola selecionada para a amostra pode ser vista como um conglomerado devido à homogeneidade entre as escolas e a heterogeneidade dos elementos que as constituem. Suas respectivas salas de aula são então vistas como outros conglomerados e uma amostra dessas é então selecionada. Só então que os alunos de cada sala são sorteados. Dessa forma, o primeiro estágio é feito a partir da seleção das escolas, o segundo da seleção das salas de aula e o terceiro e último do sorteio dos alunos de cada sala.

De maneira geral, para uma população contendo M unidades (primeiro estágio), cada uma com N subunidades (segundo estágio), cada uma com K subunidades (terceiro estágio), tem-se os valores amostrais correspondentes m, n e k, respectivamente, (Cochran, 1977). Baseado nisso, o autor define as fórmulas para a média populacional do primeiro, segundo e terceiro estágios descritas na ordem a seguir.

$$\bar{X}_{ij} = \frac{\sum_u^K x_{iju}}{K} \tag{2.14}$$

$$\bar{\bar{X}}_i = \frac{\sum_j^N \sum_u^K x_{iju}}{NK} \tag{2.15}$$

$$\bar{\bar{\bar{X}}} = \frac{\sum_i^M \sum_j^N \sum_u^K x_{iju}}{NMK} \tag{2.16}$$

onde x_{iju} é o valor obtido para a u-ésima unidade do terceiro estágio na j-ésima unidade no segundo estágio obtido da i-ésima unidade do primeiro estágio (Cochran, 1977).

Cochran (1977) define ainda as variâncias populacionais descritas como

$$S_1{}^2 = \frac{\sum_i^M (\bar{\bar{X}}_i - \bar{\bar{\bar{X}}})^2}{M-1} \tag{2.17}$$

$$S_2{}^2 = \frac{\sum_i^M \sum_j^N (\bar{X}_{ij} - \bar{\bar{X}}_{ij})}{M(N-1)} \tag{2.18}$$

$$S_3{}^2 = \frac{\sum_i^M \sum_j^N \sum_u^K (x_{ijk} - \bar{X}_{ij})}{NM(K-1)} \tag{2.19}$$

onde $S_1{}^2$, $S_2{}^2$ e $S_3{}^2$ representam as variâncias no primeiro, segundo e terceiro estágio respectivamente.

Dessa forma, o autor afirma que quando se tem amostras aleatórias em todos os estágios, a média amostral $\bar{\bar{\bar{x}}}$ por unidade do terceiro estágio é um estimador não-viesado de $\bar{\bar{\bar{X}}}$, cuja variância é

$$Var(\bar{\bar{\bar{x}}}) = \frac{1-f_1}{m} S_1{}^2 + \frac{1-f_2}{mn} S_2{}^2 + \frac{1-f_3}{mnk} S_3{}^2 \tag{2.20}$$

onde $f_1 = m/M$, $f_2 = n/N$ e $f_3 = k/K$ são as proporções amostrais nos três estágios (Cochran, 1977).

Um estimador não viesado dessa variância é dado pela fórmula a seguir (Cochran, 1977).

$$\widehat{Var}(\bar{\bar{\bar{x}}}) = \frac{1-f_1}{m} s_1{}^2 + \frac{f_1(1-f_2)}{mn} s_2{}^2 + \frac{f_1 f_2 (1-f_3)}{mnk} s_3{}^2 \tag{2.21}$$

onde s_1, s_2 e s_3 são os valores amostrais análogos a $S_1{}^2$, $S_2{}^2$ e $S_3{}^2$, respectivamente (Cochran, 1977).

Capítulo 3

PNAD

3.1 INTRODUÇÃO

A PNAD é uma pesquisa realizada anualmente, exceto nos anos de censo demográfico, em todo o país cuja população alvo é composta pelos domicílios e suas pessoas residentes na área de abrangência da pesquisa, não abrangendo, entretanto, a área rural da Região Norte (Silva et al., 2003). Seu método de coleta é amostragem probabilística de domicílios e tem plano amostral estratificado e conglomerado com até três estágios de seleção, dependendo do estrato (Silva et al., 2003).

Os estratos, também chamados estratos naturais, são escolhidos com base em 27 Unidades de Federação, pois duas ficam de fora da pesquisa por estarem inseridas na área rural da Região Norte.

3.2 Plano Amostral da PNAD

O plano amostral adotado pela PNAD é estratificado e conglomerado com um, dois ou três estágios de seleção, dependendo do estrato (Silva et al., 2003). Para a estratificação, sua amostra básica utiliza-se de duas etapas. A primeira dividindo

o país geograficamente em 36 estratos "naturais". Nessa divisão, dezoito unidades de federação formam cada uma um estrato diferente para fins de amostragem (Silva et al., 2003). As outras nove unidades de federação são subdivididas em dezoito estratos pois em cada uma foram definidos dois estratos naturais. São eles os municípios da Região Metropolitana sediada na capital do Estado e o outro os demais municípios da federação. As Notas Metodológicas da PNAD definem essas Regiões como sendo Belém, Fortaleza, Recife, Salvador, Belo Horizonte, Rio de Janeiro, São Paulo, Curitiba e Porto Alegre.

Utiliza-se de amostragem sistemática com probabilidades proporcionais ao tamanho (PPT) para a seleção dos municípios de cada unidade de federação. Os municípios pertencentes à região metropolitana são setores censitários denominados unidades primárias de amostragem (UPA). Já os domicílios seriam as unidades secundárias de amostragem (USA). É através dessa pesquisa que se mantém atualizados dados populacionais inferidos tais como renda familiar, entre outros.

Os municípios não situados na Região Metropolitana são denominados não auto-representativos (Silva et al., 2003). Esses municípios são estratificados por tamanho e proximidade geográfica, de modo a formar estratos com população total aproximadamente igual conforme os dados do último Censo (Silva et al., 2003).

A Figura 3.1 mostra claramente como se dá a divisão da Unidade de Federação. A área 1 é composta pelos municipios da região metropolitana, os quais são classificados como estratos, tendo como unidade primária cada setor e secundária cada domicilio. Já a área 2 é composta pelos municípios auto-representativos, os quais também são

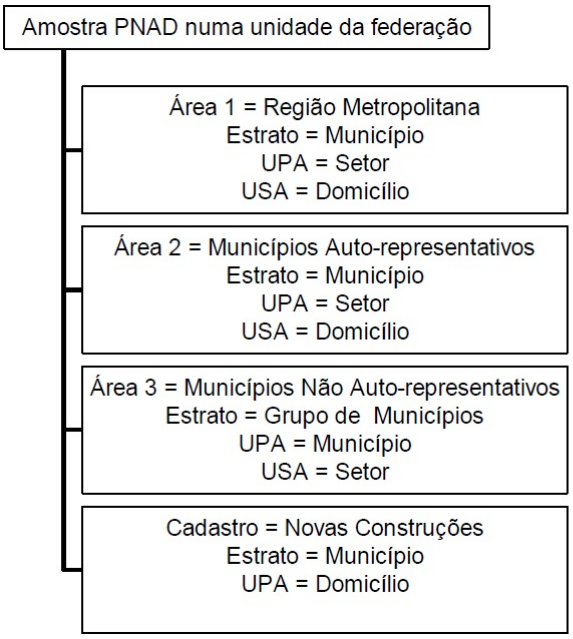

Figura 3.1: Amostra da PNAD
fonte: (Silva et al., 2003).

classificados como estratos, com unidades primária e secundária no mesmo formato da área 1. A área 3 é composta pelos municípios não auto-representativos, o estrato é formado por grupo de municípios com unidade primária sendo o domicílio e a secundária sendo o setor. Por último, tem-se o cadastro formado pelas novas construções da Unidade, sendo cada município um estrato e unidade primária o domicílio.

3.2.1 Peso

A incorporação dos pesos na estimação serve somente para a análise de medidas descritivas e pode ser feita com simplicidade empregando as opções de ponderação disponíveis nos pacotes e sistemas estatísticos padrões, tais como SAS e muitos outros (Silva et al., 2003). Já estimar medidas de dispersão, função de distribuição empírica e quantis requer considerar diversos aspectos adicionais do planejamento

da amostra usada para obter dados além dos pesos usualmente disponíveis.

O peso adotado no plano amostral da PNAD baseia-se nas informações obtidas no último censo demográfico. O desenho amostral dessa pesquisa contém todos os aspectos de um plano amostral complexo, tais como estratificação das unidades de amostragem, conglomeração (seleção da amostra em vários estágios, com unidades compostas de amostragem), probabilidades desiguais de seleção em um ou mais estágios, e ajustes dos pesos amostrais para calibração com totais populacionais conhecidos (Silva et al., 2003).

O primeiro estimador a ser calculado de alguma característica de interesse é o total, descrito na fórmula a seguir.

$$\hat{Y} = \sum_{h=1}^{H} \sum_{i=1}^{n_h} \sum_{j=1}^{n_{hi}} d_{hij} y_{hij} \tag{3.1}$$

onde H é o número de estratos existentes no estrato natural, n_h é o número de unidades primárias de amostragem, n_{hi} é o número de unidades elementares de interesse como domicílios ou pessoas pesquisadas na amostra da i-ésima UPA do h-ésimo estrato, d_{hij} é o peso amostral básico da j-ésima unidade elementar i-ésima UPA do h-ésimo estrato, y_{hij} é o valor observado da variável de interesse cujo total se deseja estimar (Silva et al., 2003).

Os pesos correspondem ao inverso das probabilidades de seleção de cada unidade ou domicílio, variam portanto dependendo do estrato natural a que pertence a unidade pesquisada (Silva et al., 2003).

O estimador $\hat{Y}$ é não-viesado para o total populacional Y no estrato natural. Visando melhorar o estimador através da incorporação de ajustes de calibração que

aproveitam informações populacionais auxiliares disponíveis, na PNAD, utiliza-se estimadores de razão os quais consideram como informação auxiliar as projeções independentes da população total para cada um dos 36 estratos (Silva et al., 2003).

Dessa forma, o estimador razão empregado é definido como

$$\hat{Y}_R = \hat{Y}\frac{P}{\hat{P}} = P\hat{R} \tag{3.2}$$

onde P é a população residente projetada para o estrato natural, $\hat{P}$ seu respectivo estimador, ou seja, $\hat{P} = \sum_{h=1}^{H}\sum_{i=1}^{n_h}\sum_{j=1}^{n_{hi}} d_{hij}x_{hij}$ onde x_{hij} é o número de moradores do j-ésimo domicílio pesquisado na i-ésima UPA do h-ésimo estrato (Silva et al., 2003).

Sendo assim, cada unidade amostrada tem um peso ajustado, o qual é calculado e adicionado aos registros de dados da PNAD. Esse peso é dado por (Silva et al., 2003):

$$w_{hij} = d_{hij}\frac{P}{\hat{P}} \tag{3.3}$$

O estimador do total passa a ser então calculado por:

$$\hat{Y}_R = \sum_{h=1}^{H}\sum_{i=1}^{n_h}\sum_{j=1}^{n_{hi}} w_{hij}y_{hij} \tag{3.4}$$

Os pesos assim ajustados, quando usados para estimar o total da população em cada estrato natural, produzem uma estimativa que é igual ao valor da população residente projetada para o estrato natural pelo IBGE, conferindo assim a propriedade de calibração no total populacional à amostra da PNAD (Silva et al., 2003).

Capítulo 4

MODELO DE REGRESSÃO UTILIZANDO PLANO AMOSTRAL

Existem várias metodologias sugeridas para incorporar o plano amostral na análise de regressão. O modelo de regressão especificado por Pessoa and Silva (1998), por exemplo, utiliza uma aplicação do Método de Máxima Pseudo-Verossimilhança para considerar o plano amostral. Porém, para a estimação pontual dos coeficientes, basta incluir os pesos amostrais no cálculo, o que faz com que a estimativa seja a mesma obtida por máxima verossimilhança ou muito próxima, se não a mesma, de outros métodos também utilizados.

Dessa forma, os métodos de destaque no presente trabalho são aqueles utilizados no cálculo da variância dos estimadores. A maneira como de estimar a variância para dados amostrais complexos é significativamente diferente em comparação ao cálculo para amostra aleatória simples.

4.1 Amostragem Aleatória Simples

Para o caso de levantamento de dados via amostragem aleatória simples, tem-se

o modelo clássico de regressão múltipla descrito por (Kutner et al., 2004). Esse modelo, o qual considera o caso geral de p variáveis explicativas pode ser descrito de acordo com:

$$Y_i = \beta_0 + \sum_{j=1}^{p} \beta_j X_{ij} + \epsilon_i \tag{4.1}$$

onde $i = 1, ..., n$, sendo n o número total de observações.

Na Equação (4.1), para cada i-ésima observação, X_{ij} corresponde a uma variável que supostamente explica o modelo, fazendo com que Y_i seja a variável dependente de interesse, também conhecida como variável resposta. Os coeficientes β_j, $j = 1, ..., p$, são conhecidos como parâmetros da regressão. A variável ϵ_i corresponde ao erro aleatório associado à i-ésima observação no modelo. Kutner et al. (2004) define esse modelo como um hiperplano, ou seja, um plano contido num espaço p-dimensional. O modelo segue os pressupostos de que X_{ij} são constantes e conhecidos e que ϵ_i são variáveis aleatórias independentes, se distribuem normalmente com média igual a zero e variância constante. Sendo assim,

$$E(Y_i) = \beta_0 + \beta_1 X_{i1} + ... + \beta_p X_{ip} \tag{4.2}$$

com $E(Y_i)$ correspondendo ao valor médio da variável resposta.

Vale notar que cada coeficiente β_j determina a mudança média na variável resposta Y_i quando o valor de X_{ij} varia. Especificamente, o coeficiente β_0 representa o valor médio da variável resposta quando $X_{ij} = 0$, para todo $j = 1, ..., p$. Por isso esse coeficiente é chamado de intercepto da regressão.

Para a visualização dos dados como um todo, a Equação (4.1) é comumente

manipulada usando a sua definição matricial:

$$\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon} \tag{4.3}$$

onde $\boldsymbol{Y}$ é o vetor $n \times 1$ dos valores observados da variável resposta, também denominado vetor observacional, $\boldsymbol{\beta}$ é o vetor $(p+1) \times 1$ dos coeficientes da regressão, $\boldsymbol{\epsilon}$ o vetor $n \times 1$ dos resíduos ou erros da regressão e $\boldsymbol{X}$ é a matriz $n \times p$ da forma:

$$\boldsymbol{X} = \begin{bmatrix} 1 & X_{11} & ... & X_{1p} \\ 1 & X_{21} & ... & X_{2p} \\ : & : & : & : \\ 1 & X_{n1} & ... & X_{np} \end{bmatrix}$$

Assim, a média do vetor resposta $\boldsymbol{Y}$ fica:

$$E(\boldsymbol{Y}) = E(\boldsymbol{X}\hat{\boldsymbol{\beta}}) = \boldsymbol{X}\boldsymbol{\beta} \tag{4.4}$$

onde $\hat{\boldsymbol{\beta}}$ é o estimador do vetor $\boldsymbol{\beta}$. A Equação (4.3) é conhecida como Modelo Geral de Regressão Linear (Kutner et al., 2004).

Vale notar também que da Equação (4.3) tem-se:

$$Var(\boldsymbol{Y}) = Var(\boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}) = Var(\boldsymbol{\epsilon}) \tag{4.5}$$

Pois $\boldsymbol{X}\boldsymbol{\beta}$ é constante, logo, tem variância igual a zero.

4.1.1 Estimação dos coeficientes de regressão

Para estimar os coeficientes β_j da regressão, estima-se diretamente o vetor $\boldsymbol{\beta}$ utilizando-se o método de Mínimos Quadrados ou o Método da Máxima Verossimilhança, ambos definidos em (Kutner et al., 2004). O método dos mínimos quadrados consiste em considerar o desvio de cada Y_i do seu respectivo valor esperado, como mostrado a seguir.

$$Y_i - (\beta_0 + \sum_{j=1}^{p} \beta_j X_j) \tag{4.6}$$

Por ser um método geral, ou seja, utilizado para uma amostra com n observações, é preciso considerar o desvio total de todos os n desvios existentes. Como cada um pode ser positivo ou negativo, simplesmente somar esses desvios poderia anular o desvio total, prejudicando a análise. Dessa forma, o que se faz então é calcular a soma quadrática desses desvios, com isso a fórmula de mínimos quadrados fica:

$$Q = \sum_{i=1}^{n} [Y_i - (\beta_0 + \sum_{j=1}^{p} \beta_j X_j)]^2 \tag{4.7}$$

De acordo com o método de Mínimos Quadrados, os estimadores $\hat{\beta}_j$ dos coeficientes são aqueles que minimizam o critério Q descrito na Equação (4.7). Para encontrá-los, primeiramente, define-se os estimadores $\hat{\beta}_j$ por meio de um vetor $\hat{\boldsymbol{\beta}}$ dado por:

$$\hat{\boldsymbol{\beta}} = \begin{bmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \\ \vdots \\ \hat{\beta}_p \end{bmatrix}$$

A forma matricial desse método é descrita deduzindo-se da Equação (4.4).

$$\hat{\boldsymbol{Y}} = \boldsymbol{X}\hat{\boldsymbol{\beta}} \tag{4.8}$$

Assim, utilizando manipulações algébricas descritas em Kutner et al. (2004), os estimadores de mínimos quadrados são:

$$\hat{\boldsymbol{\beta}} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{Y} \tag{4.9}$$

no qual $\boldsymbol{X}'$ representa a matriz transposta de $\boldsymbol{X}$ e $(\boldsymbol{X}'\boldsymbol{X})^{-1}$ é uma matriz simétrica, inversa da matriz $(\boldsymbol{X}'\boldsymbol{X})$.

A variância desse estimador é calculada como:

$$Var[\hat{\boldsymbol{\beta}}] = Var[(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{Y}] \tag{4.10}$$

Após algumas manipulações algébricas com base em propriedades de cálculo de variância, tem-se que:

$$Var[\hat{\boldsymbol{\beta}}] = (\boldsymbol{X}'\boldsymbol{X})^{-1}Var[\boldsymbol{Y}] \tag{4.11}$$

4.1.2 Estimação da variância dos resíduos

Considerando que os resíduos ϵ_i são variáveis aleatórias, o modelo pressupõe que sua variância seja constante e igual a σ^2. Para estimar o seu valor utiliza-se os mesmos princípios do método de mínimos quadrados.

$$\boldsymbol{Y} - \hat{\boldsymbol{Y}} = \hat{\boldsymbol{\epsilon}} \tag{4.12}$$

De tal forma, que a soma de quadrados é dada por

$$\hat{\boldsymbol{\epsilon}}'\hat{\boldsymbol{\epsilon}} = [\boldsymbol{Y} - \hat{\boldsymbol{Y}}]'[\boldsymbol{Y} - \hat{\boldsymbol{Y}}] \tag{4.13}$$

comumente chamada de *SSE* (*Sum of Squared Errors*), sigla em inglês para Soma de Quadrados dos resíduos, onde $[\boldsymbol{Y} - \hat{\boldsymbol{Y}}]'$ é a matriz transposta da matriz $[\boldsymbol{Y} - \hat{\boldsymbol{Y}}]$ (Kutner et al., 2004).

Por (4.9), tem-se então:

$$\hat{\boldsymbol{\epsilon}}'\hat{\boldsymbol{\epsilon}} = [\boldsymbol{Y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}]'[\boldsymbol{Y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}] \tag{4.14}$$

Em geral, procura-se definir essa soma de quadrados em função dos parâmetros conhecidos no modelo, no caso a matriz $\boldsymbol{X}$. Baseado na Equação (4.8), tem-se:

$$\hat{\boldsymbol{\epsilon}}'\hat{\boldsymbol{\epsilon}} = [\boldsymbol{Y} - \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{Y}]'[\boldsymbol{Y} - \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{Y}] \tag{4.15}$$

Definindo a matriz $\boldsymbol{H} = \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'$ e isolando o vetor $\boldsymbol{Y}$ na equação, chega-se ao resultado:

$$\hat{\boldsymbol{\epsilon}}'\hat{\boldsymbol{\epsilon}} = \boldsymbol{Y}'[\boldsymbol{I} - \boldsymbol{H}]'[\boldsymbol{I} - \boldsymbol{H}]\boldsymbol{Y} \tag{4.16}$$

Essa soma possui $n - p$ graus de liberdade, pois p graus são perdidos quando estima-se os p parâmetros da regressão. Assim sendo, tem-se o estimador conhecido como Quadrado Médio do Erro, sigla em inglês *MSE* (*Mean Squared Errors*) definido por:

$$MSE = \frac{SSE}{n - p} \tag{4.17}$$

De forma que $MSE = \hat{\sigma}^2$, pois

$$E[MSE] = \sigma^2$$

Sendo assim, tem-se que a variância do vetor residual $\boldsymbol{\epsilon}$ pode ser definida como:

$$Var(\boldsymbol{\epsilon}) = \boldsymbol{I}\sigma^2 \tag{4.18}$$

onde $\boldsymbol{I}$ é a matriz identidade $n \times n$.

Assim, da Equação (4.5) tem-se:

$$Var(\boldsymbol{Y}) = \boldsymbol{I}\sigma^2 \tag{4.19}$$

4.2 Amostragem Complexa

Para dados provenientes de amostragem complexa, outras formas de cálculo para estimação dos coeficientes e variância de modelos de regressão são adotadas. As técnicas para a estimação pontual dos parâmetros são mais gerais, pois a fórmula permanece a mesma (e assim o seu resultado) independente do plano amostral utilizado, diferenciando-se apenas por incluir os pesos amostrais. Já para o cálculo da variância, apresenta-se o algoritmo iterativo mais adotado.

4.2.1 Estimação dos coeficientes de regressão

O cálculo para estimar os coeficientes da regressão clássica é feito utilizando-se o método dos mínimos quadrados ponderados. Esse método é análogo ao mostrado na Seção 4.2.1, diferenciando-se por incorporar os pesos amostrais. Dessa forma, os coeficientes da regressão para dados complexos é calculado por (SAS, 2012):

$$\hat{\boldsymbol{\beta}} = (\boldsymbol{X}'\boldsymbol{W}\boldsymbol{X})^{-}\boldsymbol{X}'\boldsymbol{W}\boldsymbol{Y} \tag{4.20}$$

em que $(\boldsymbol{X}'\boldsymbol{W}\boldsymbol{X})^{-}$ é a inversa generalizada da matriz $\boldsymbol{X}'\boldsymbol{W}\boldsymbol{X}$.

4.2.2 Estimação da variância dos resíduos

No caso da variância do estimador $\hat{\boldsymbol{\beta}}$, o procedimento SURVEYREG adota quatro metodologias diferentes para o seu cálculo. São elas a linearização por séries de Taylor, replicação repetida balanceada (*BRR* em inglês), uma modificação de *BRR* conhecida como *Fay's BRR method* e *jackknife* (SAS, 2012). Aqui é descrito o método de linearização por Taylor, o qual é o padrão do SURVEYREG. De acordo com SAS (2012), esse é o método mais comumente usado para estimar a matriz de variâncias e covariâncias dos coeficientes de regressão no caso de amostragem complexa. Seja $\boldsymbol{r} = \boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}$, tem-se a variância de $\hat{\boldsymbol{\beta}}$ dada por:

$$\widehat{\boldsymbol{Var}}(\hat{\boldsymbol{\beta}}) = (\boldsymbol{X}'\boldsymbol{W}\boldsymbol{X})^{-}\boldsymbol{G}(\boldsymbol{X}'\boldsymbol{W}\boldsymbol{X})^{-} \tag{4.21}$$

A matriz $\boldsymbol{G}$ da Equação (4.21) é definida como:

$$\boldsymbol{G} = \frac{n-1}{n-p}\sum_{h=1}^{H}\frac{n_h(1-f_h)}{n_h-1}\sum_{i=1}^{n_h}(\boldsymbol{e_{hi.}} - \bar{\boldsymbol{e}}_{\boldsymbol{h..}})'(\boldsymbol{e_{hi.}} - \bar{\boldsymbol{e}}_{\boldsymbol{h..}}) \tag{4.22}$$

onde $\boldsymbol{e}_{hij} = w_{hij} r_{hij} \boldsymbol{x}_{hij}$, $\boldsymbol{e}_{hi.} = \sum_{j=1}^{m_{hi}} \boldsymbol{e}_{hi.}$, $\bar{\boldsymbol{e}}_{h..} = \frac{1}{n_h} \sum_{i=1}^{n_h} \boldsymbol{e}_{hi}$, para os quais w_{hij} é o peso associado a cada observação, H é o número de estratos, n_h é o número de conglomerados dentro do h-ésimo estrato e m_{hi} é o número de elementos dentro do i-ésimo conglomerdo contido no h-ésimo estrato.

Capítulo 5

MATERIAL E MÉTODOS

5.1 Material

O material a ser utilizado no estudo consiste em dados da PNAD de 2001 até 2011. Para um melhor aproveitamento e resumo de registros a serem observados e analisados, utilizou-se apenas os bancos de dados dos anos ímpares.

Foi observado que as variáveis de identificação para *cluster* e estrato e as variáveis correspondentes aos pesos utilizados na amostragem não mudam sua nomenclatura com o passar do tempo, exceto para o ano 2003 em que a variável de identificação do município é denominada v4615. Como apenas essa variável sofre mudança para apenas esse ano, isso pode ser alterado no próprio programa, renomeando a variável. Em outras palavras, de acordo com o dicionário da PNAD, as variáveis a serem utilizadas no programa permanecem as mesmas de 2001 a 2011. A Tabela 5.1 mostra exatamente isso.

Tendo isso em mente, é possível utilizar a mesma programação para analisar os dados da PNAD durante os últimos dez anos e verificar se existe diferença nas estimativas pontuais dos coeficientes e em seus respectivos erros padrão. A metodologia utilizada nessa análise é descrita na seção 5.2.

Tabela 5.1: Variáveis Correspondentes - PNAD

Variável Correspondente	**2001**	**2003*** **	**2005**	**2007**	**2009**	**2011**
probabilidade do município*	v4605	v4605	v4605	v4605	v4605	v4605
probabilidade do setor censitário**	v4607	v4607	v4607	v4607	v4607	v4607
peso do domicílio	v4611	v4611	v4611	v4611	v4611	v4611
peso da pessoa	v4729	v4729	v4729	v4729	v4729	v4729
identificação do estrato	v4602	v4602	v4602	v4602	v4602	v4602
identificação do município	UPA	v4615	UPA	UPA	UPA	UPA
identificação do setor censitário	v0102	v0102	v0102	v0102	v0102	v0102

* peso do município=1/(probabilidade do município)
** peso do setor=1/(probabilidade do setor)
***único ano para o qual a variável UPA é diferente

5.2 Métodos

Partiu-se da ideia de uma regressão onde a variável dependente utilizada seria REMUNERAÇÃO e a variável independente ANOS DE ESTUDO. A variável dependente foi criada a partir do somatório de todas as variáveis relativas a renda de cada pessoa dentro do domicílio. Os dados utilizados são provenientes das PNADs dos anos ímpares de 2001 até 2011, pela junção das tabelas PESSOAS e DOMICÍLIOS dos anos utilizados.

Primeiramente foi feita uma regressão clássica, ou seja, considerando população infinita e amostragem aleatória simples, utilizando o procedimento PROC REG sem peso. Em seguida foi feito o mesmo procedimento mas levando em conta os pesos das observações no modelo. Após isso, o procedimento executado foi o PROC SURVEYREG, também considerando os pesos das observações. Depois, repetiu-se o procedimento anterior considerando não só os pesos da população mas também incorporando a estratificação e a conglomeração em um estágio. Por último, fez-se

uso da PROC SURVEYREG para incorporar o plano amostral adotado pela PNAD.

Esses procedimentos tiveram como objetivo comparar as estimativas dos coeficientes da regressão e seus respectivos erros padrão e verificar se existe diferença significativa nesses valores dada a análise com ou sem o plano amostral incorporado. O último procedimento foi feito três vezes tendo em vista que, para o correto cálculo da variância, é preciso executar uma PROC SURVEY para cada estágio de seleção pois os dados da PNAD são coletados em três estágios.

A programação realizada encontra-se a seguir.

```
/*REGRESSÃO CLÁSSICA SEM PESO*/
proc reg data=pes2007;
 model remuneracao=anos_estudo;
run;
quit;

/*REGRESSÃO CLÁSSICA COM PESO*/
proc reg data=pes2007;
 model remuneracao=anos_estudo;
 weight v4729;
run;
quit;

/*REGRESSÃO SURVEY COM PESO*/
proc surveyreg data=pes2007;
 model remuneracao=anos_estudo;
 weight v4729;
run;
quit;

/*REGRESSÃO SURVEY COM PESO E FPC*/
data pes2007;set pes2007;_rate_=1/v4611;run;
proc surveyreg data=pes2007 rate=pes2007;
 model remuneracao=anos_estudo;
 weight v4729;
run;
quit;

/********************** regressão dados complexos ******/
/*PRIMEIRO ESTÁGIO*/
data pes2007;set pes2007;_rate_1=v4605;run;
ods output parameterestimates=beta1;
```

```
proc surveyreg data=pes2007 rate=pes2007 (rename=_rate_1=_rate_);
 model remuneracao=anos_estudo;
 strata v4602 /nocollapse;
 cluster UPA; *cluster v4618;
 weight v4729;
run;
quit;

/*SEGUNDO ESTÁGIO*/
data pes2007;set pes2007;
 _rate2_=1-v4605*(1-v4607);
 if _rate2_=. then delete;
run;
ods output parameterestimates=beta2;
proc surveyreg data=pes2007 rate=pes2007(rename=_rate2_=_rate_);
 model remuneracao=anos_estudo;
 strata UPA;*strata v4618;
 weight v4729;
run;
quit;

/*TERCEIRO ESTÁGIO*/
data pes2007;set pes2007;
 _rate3_=1-v4605*v4607*(1-1/v4611);
 if _rate3_=. then delete;
run;
ods output parameterestimates=beta3;
proc surveyreg data=pes2007 rate=pes2007(rename=_rate3_=_rate_);
 model remuneracao=anos_estudo;
 strata v0102;
 weight v4729;
run;
quit;

data beta;merge beta1(rename=(StdErr=StdErr1 tValue=tValue1))
 beta2(rename=(StdErr=StdErr2 tValue=tValue2))
 beta3(rename=(StdErr=StdErr3 tValue=tValue3));
 StdErr=sqrt(StdErr1**2+StdErr2**2+StdErr3**2);
 tValue=estimate/StdErr;
run;
```

Capítulo 6

ANÁLISE DOS RESULTADOS

A sequência de procedimentos descritos na última sessão foram executados, apenas para exemplo, utilizando os dados do ano de 2007, podendo ser estendida para dados da PNAD de outros anos, pois, como foi visto anteriormente, as variáveis utilizadas na programação para as estimativas pontuais e cálculo do erro padrão permanecem as mesmas de 2001 até o ano de 2011. A programação foi utilizada com o intuito de comparar os cinco métodos diferentes de análise dos dados.

O procedimento de inclusão do plano amostral visa calcular a variância dada pela Equação (2.21). Cada PROC SURVEY fornece a variância de um estágio de seleção, sendo necessário somar as três variâncias obtidas para o cálculo da variância total estimada. Isso é feito definindo cada *output* como um *data* e, no final, juntando esses valores em um só.

Visando comparar os resultados obtidos por meio dos diferentes procedimentos utilizados no SAS, foi criada uma tabela mostrando esses resultados para cada procedimento executado. As estimativas dos coeficientes e seus respectivos erros padrão obtidos por cada PROC do programa referido foram indexadas na Tabela 6.1. Os resultados obtidos para o último procedimento são calculados como dois estágios,

conforme explicado mais adiante.

Tabela 6.1: Estimativa dos coeficientes

PROCEDIMENTO	**COEFICIENTE ESTIMADO**	**ESTIMATIVA PONTUAL**	**ERRO PADRÃO**
PROC REG	INTERCEPTO	6,2230100	1,50682000
(sem os pesos)	ANOS DE ESTUDO	101,7026400	1,31954000
PROC REG	INTERCEPTO	52,5637100	9,64473000
(com os pesos)	ANOS DE ESTUDO	96,9074800	1,22296000
PROC SURVEYREG	INTERCEPTO	52,5637149	9,64900000
(com pesos)	ANOS DE ESTUDO	96,9074824	1,94282345
PROC SURVEYREG	INTERCEPTO	52,5637149	9,21411450
(com estrato e *cluster*)	ANOS DE ESTUDO	96,9074824	1,85521335
REGRESSÃO PNAD	INTERCEPTO	52,5637149	10,1303
(complexo)	ANOS DE ESTUDO	96,9074824	2,0600

De acordo com a Tabela 6.1 é possível notar a influência que o plano amostral e os pesos exercem sobre as estimativas. Para o primeiro caso, em que foi utilizado o procedimento clássico de regressão, as estimativas pontuais são completamente diferentes das obtidas pelos outros procedimentos, assim como o erro padrão de cada uma. Do segundo procedimento em diante as estimativas pontuais mostram-se muito próximas para cada coeficiente, porém vale notar que o erro padrão sofre mudanças um pouco mais significativas, de $1,22$ para $1,94$ e $1,86$, nos três procedimentos para o coeficiente da variável independente. Já quando se incorporam os três estágios de seleção no cálculo da variância, o erro padrão do intercepto e do coeficiente da variável independente são aproximadamente $5,32$ e $1,07$, respectivamente. Tais resultados mostram que para o cálculo da variância total o plano amostral deve ser incorporado na análise pois este é um fator determinante na sua estimativa.

A fim de verificar o quanto cada estágio de seleção influencia no cálculo da variância total, observou-se os respectivos erros padrão obtidos em cada SURVEY-

REG, com e sem o fator de correção populacinal de cada estágio (*fpc*). Como a variável que identifica o município selecionado corresponde ao (*cluster*) no primeiro estágio e ao estrato no segundo, foram feitas análises utilizando, primeiro, a variável UPA e depois a variável v4618. Os resultados obtidos, utilizando a primeira variável, pela análise de regressão do primeiro, segundo e terceiro estágio estão disponibilizados, respectivamente, nas Tabelas 6.2 até 6.7 e nas Tabelas 6.8 até 6.11.

Tabela 6.2: Estimativa dos coeficientes no primeiro estágio (com *fpc*) - UPA

COEFICIENTE	**ESTIMATIVA PONTUAL**	**ERRO PADRÃO**
INTERCEPTO	52,5637149	5,32327415
ANOS DE ESTUDO	96,9074824	1,06817261

Tabela 6.3: Estimativa dos coeficientes no primeiro estágio (sem *fpc*) - UPA

COEFICIENTE	**ESTIMATIVA PONTUAL**	**ERRO PADRÃO**
INTERCEPTO	52,5637149	5,68796021
ANOS DE ESTUDO	96,9074824	1,15062836

Tabela 6.4: Estimativa dos coeficientes no segundo estágio (com *fpc*) - variável UPA

COEFICIENTE	**ESTIMATIVA PONTUAL**	**ERRO PADRÃO**
INTERCEPTO	52,5637149	8,61886177
ANOS DE ESTUDO	96,9074824	1,76140203

Tabela 6.5: Estimativa dos coeficientes no segundo estágio (sem *fpc*) - variável UPA

COEFICIENTE	**ESTIMATIVA PONTUAL**	**ERRO PADRÃO**
INTERCEPTO	52,5637149	9,57390960
ANOS DE ESTUDO	96,9074824	1,92436584

Nas Tabelas 6.6 e 6.7 é interessante notar que ao ignorar o fator de correção populacional no terceiro estágio, o erro padrão, apesar de pequeno comparado aos

Tabela 6.6: Estimativa dos coeficientes no terceiro estágio (com *fpc*) - variável UPA

COEFICIENTE	**ESTIMATIVA PONTUAL**	**ERRO PADRÃO**
INTERCEPTO	52,5637149	0,00085573
ANOS DE ESTUDO	96,9074824	0,00007806

Tabela 6.7: Estimativa dos coeficientes no terceiro estágio (sem *fpc*) - variável UPA

COEFICIENTE	**ESTIMATIVA PONTUAL**	**ERRO PADRÃO**
INTERCEPTO	52,5637149	0,01131773
ANOS DE ESTUDO	96,9074824	0,00102921

outros estágios, é maior que zero a partir da segunda e terceira casa decimal. Em outras palavras, ao incluir o *fpc* no terceiro estágio a variância aumenta. O mesmo ocorre nos outros estágios de seleção, indicando assim que a variância total aumenta significativamente ao ignorar o fator de correção.

Utilizando a variável v4618 como identificação dos municípios há mudança nos resultados. As estimativas pontuais e erro padrão obtidos em cada estágio são mostradas nas tabelas a seguir.

Tabela 6.8: Estimativa dos coeficientes no primeiro estágio (com *fpc*) - variável v4618

COEFICIENTE	**ESTIMATIVA PONTUAL**	**ERRO PADRÃO**
INTERCEPTO	52,5637149	5,35852627
ANOS DE ESTUDO	96,9074824	1,05707613

Tabela 6.9: Estimativa dos coeficientes no primeiro estágio (sem *fpc*) - variável v4618

COEFICIENTE	**ESTIMATIVA PONTUAL**	**ERRO PADRÃO**
INTERCEPTO	52,5637149	12,6027111
ANOS DE ESTUDO	96,9074824	2,7010110

A análise no terceiro estágio não foi repetida para essa variável pois nessa fase de coleta não há influência de município, tendo em vista que no último estágio são

Tabela 6.10: Estimativa dos coeficientes no segundo estágio (com *fpc*) - variável v4618

COEFICIENTE	**ESTIMATIVA PONTUAL**	**ERRO PADRÃO**
INTERCEPTO	52,5637149	7.99520928
ANOS DE ESTUDO	96,9074824	1,62037810

Tabela 6.11: Estimativa dos coeficientes no segundo estágio (sem *fpc*) - variável v4618

COEFICIENTE	**ESTIMATIVA PONTUAL**	**ERRO PADRÃO**
INTERCEPTO	52,5637149	9,30538496
ANOS DE ESTUDO	96,9074824	1,87955952

selecionados os domicílios dentro de cada setor. Vale observar também que usando a variável v4618 o erro padrão muda significativamente ao ignorar o *fpc*. Já utilizando a variável UPA essa diferença, para cada estágio, é muito pequena, mostrando assim que a análise é mais robusta com esta última variável.

Pelo exposto nota-se que, com o fator de correção, o erro padrão, tanto para o intercepto quanto para o coeficiente da variável independente, aumenta do primeiro para o segundo estágio e diminui do segundo para o terceiro. Observa-se também que, na Tabela 6.7, o erro padrão para cada coeficiente é maior que zero somente a partir da quarta e quinta casa decimal. Isso mostra que para o cálculo da variância total, o terceiro estágio não tem muita influência, indicando assim que a sua omissão não traria prejuízo ao resultado. Com isso, os resultados obtidos incorporando os dois estágios são expostos nas Tabelas 6.12 e 6.13, onde cada erro padrão é calculado como a raíz quadrada da soma das variâncias de cada estágio.

Dessa forma, o objetivo do trabalho foi justamente obter um algoritmo que calculasse a variância total em apenas um procedimento. Assim, o usuário que necessite

Tabela 6.12: Estimativa final dos coeficientes (com *fpc*)

COEFICIENTE	**ESTIMATIVA PONTUAL**	**ERRO PADRÃO**
INTERCEPTO	52,5637149	10,1303
ANOS DE ESTUDO	96,9074824	2,0600

Tabela 6.13: Estimativa final dos coeficientes (sem *fpc*)

COEFICIENTE	**ESTIMATIVA PONTUAL**	**ERRO PADRÃO**
INTERCEPTO	52,5637149	11,1361
ANOS DE ESTUDO	96,9074824	2,2421

desse tipo de análise não precisará ter o conhecimento prévio do plano amostral da PNAD, pois o mesmo já estará incorporado ao algoritmo.

O algoritmo foi elaborado dentro da PROC IML do SAS. Como foi construído com o intuito de poder ser utilizado na análise de qualquer banco de dados da PNAD, o programa foi disposto em uma *macro*, a qual pode ser executada para quaisquer dados que atendam os pressupostos amostrais.

As notas metodológicas da PNAD afirmam que o primeiro estágio de seleção abrange os municípios, logo esses poderiam ser vistos como unidades primárias de seleção. Porém, de acordo com o dicionário das variáveis, existe uma variável que define as unidades primárias (v4618) e outra que define a delimitação dos municípios (UPA). Dessa forma, foram executados procedimentos utilizando tanto a primeira variável como a segunda com o objetivo de comparar seus resultados. Para cada análise, considerou-se a inclusão ou não da taxa (*fpc*), como pôde ser visto anteriormente.

A programação está disponível no Apêndice A deste trabalho. Seus resultados podem ser visualizados nas Tabelas 6.14 e 6.15, respectivamente com e sem o fator

de correção. Sabendo que a diferença no cálculo da variância incorporando o terceiro estágio foi pouco significativo, a variância no algoritmo também foi calculada considerando-se apenas 2 estágios de seleção. A variável utilizada no algoritmo para identificar os municípios foi UPA, dada a robustez na análise com sua utilização.

Tabela 6.14: Estimativa dos coeficientes com *fpc* (algoritmo)

PROCEDIMENTO	**COEFICIENTE ESTIMADO**	**ESTIMATIVA PONTUAL**	**ERRO PADRÃO**
REGRESSÃO PNAD (complexo)	INTERCEPTO	52,5637149	10,1293
	ANOS DE ESTUDO	96,9074824	2,0598

Tabela 6.15: Estimativa dos coeficientes sem *fpc* (algoritmo)

PROCEDIMENTO	**COEFICIENTE ESTIMADO**	**ESTIMATIVA PONTUAL**	**ERRO PADRÃO**
REGRESSÃO PNAD (complexo)	INTERCEPTO	52,5637149	11,1361
	ANOS DE ESTUDO	96,9074824	2,2421

Nota-se que os resultados com o algoritmo são os mesmos que foram obtidos com os procedimentos SURVEY mostrados nas tabelas 6.12 e 6.13.

Além disso, a saída do programa ficou idêntica ao *output* da PROC SURVEYREG, isso porque no final do algoritmo existe um comando *print* pra cada um desses dados do *output*. Uma outra característica interessante é que o usuário terá a liberdade de escolher se quer fazer a sua análise considerando ou não o *fpc* e o intercepto. O método padrão do programa faz a análise considerando esses dois elementos, porém existem as opções INTERCEPT=NO e FPC=NO na chamada da *macro*, de forma que caso o usuário opte por ambas essas opções, o programa ignora esses fatores no seu cálculo, tendo em vista que eles estão dentro de condicionais, como mostrado a seguir.

```
%if %upcase(&intercept)=YES %then %do;
 x=j(n,1,1)||x;
 nomes={Intercept &x}‘;
%end;

%if %upcase(&fpc)=YES %then %do;
 G1=((n-1)/(n-ncol(x)))*((((ehi[,ncol(ehi)-1]#(1-ehi[,
 ncol(ehi)-2]))/(ehi[,
 ncol(ehi)-1]-1))#(ehi[,3:2*ncol(x)]-ehi[,
 3+ncol(x):2+2*ncol(x)]))‘*(ehi[,
 3:2*ncol(x)]-ehi[,3+ncol(x):2+2*ncol(x)]));
%end;

%if %upcase(&fpc)=YES %then %do;
 eh=e2||ehi2[,1:ncol(ehi2)-1]||1-fh1#(1-fh2)||ehi2[,ncol(ehi2)];
%end;

%if %upcase(&fpc)=YES %then %do;
 G2=((n-1)/(n-ncol(x)))*((((eh[,ncol(eh)]#(1-eh[,
 ncol(eh)-1]))/(eh[,ncol(eh)]-1))#(eh[,1:ncol(x)]-eh[,
 1+ncol(x):2*ncol(x)]))‘*(eh[,1:ncol(x)]-eh[,1+ncol(x):2*ncol(x)]));
%end;
```

A saída para o programa elaborado está ilustrada a seguir. Esta é a saída obtida para o método *default* do programa, ou seja, considerando o intercepto e o *fpc* na análise de regressão.

```
                    Design Summary

                Number of Strata            545
               Number of Clusters           817


            Number of Observations Used      57862
                   Sum of Weights         28002586
            Weighted Mean of remuneracao  706.08461
                      Root MSE           1202.7752


      Regression Analysis for Dependent Variable remuneracao


            Estimated Regression Coefficients
```

Parameter	Estimate	Std. Error	t Value	P > |t|
INTERCEPT	52.56371489	10.12933155	5.19	<.0001
ANOS_ESTUDO	96.90748239	2.05985089	47.05	<.0001

NOTE: The denominator degrees of freedom for the t tests is 272 .

Capítulo 7

CONCLUSÃO

O trabalho mostrou a importância de sempre incluir nos cálculos da análise inferencial não só os pesos das observações mas as técnicas de amostragem que foram utilizadas, pois nota-se a partir dos resultados expostos que ao incorporar os estágios de seleção com seus respectivos pesos de estrato e conglomerado há significativa alteração nos erros padrão dos coeficientes estimados. A partir dos resultados exibidos no Capítulo anterior, pode-se observar que de fato essa alteração ocorre em cada estágio. Vale notar que, a partir da inclusão dos pesos, independente de a análise ser clássica ou do tipo SURVEY, as estimativas pontuais não se alteram. Quanto ao erro padrão, o valor muda para cada procedimento.

Já com relação à *macro* elaborada, pode-se dizer que esta mostrou-se eficiente. Isso se deve ao fato de que, primeiramente, tanto as estimativas pontuais quanto os erros padrão calculados com o algoritmo foram os mesmos obtidos com a sequência de procedimentos SURVEYREG em dois estágios, além disso as fórmulas incluídas nele foram baseadas nas fórmulas pertencentes ao algoritmo interno do SAS (SAS, 2012) para tais procedimentos. Também deve-se levar em conta que o usuário que necessitar de uma análise desse tipo com dados da PNAD não precisa ter conheci-

mento algum do plano amostral ou dos pesos incluídos nos estágios de seleção, pois todos esses valores já estão incorporados ao algoritmo. Sendo assim, quem necessitar trabalhar com bancos de dados como esse encontrará muita facilidade utilizando esse programa, pois basta declarar apenas a tabela e as variáveis explicativas, visto que o intercepto e o *fpc* já estão incluídos como procedimento padrão. Caso o usuário não queira utilizá-los, basta adicionar os comandos INTERCEPT=NO e FPC=NO na chamada da *macro*, como mostrado no Apêndice.

Com isso, qualquer um que tiver acesso ao algoritmo poderá fazer análises de regressão de dados da PNAD seguro de que suas estimativas são, não apenas, não-viesadas mas com precisão confiável, já que o erro padrão calculado leva em conta todos os estágios de seleção. Com estimativa correta de coeficientes e com precisão bem calculada, o risco de considerar a variável no modelo quando na verdade esta não é significativa ou de considerá-la como não significativa, quando na verdade esta deve ser incluída são minimizados.

Referências Bibliográficas

Cochran, W. G. (1977). *Sampling Techniques*, (3rd ed.). Wiley.

de Sousa, M. H. & da Silva, N. N. (2000). Comparação de softwares para análise de dados de levantamentos complexos. *Rev Saúde Pública*, 34(6):646–53.

Faiella, I. (2010). The use of survey weights in regression analysis. *Bank of Italy Temi di Discussione (Working Paper) No*, 739.

Franco, J. V. & de Moraes, J. R. Envelhecimento populacional brasileiro: O desafio da capacidade funcional.

George, D. (2003). *SPSS for Windows Step by Step: A Simple Study Guide and Reference, 17.0 Update, 10/e*. Pearson Education India.

Kish, L. (1965). *S UR VE Y SAMPLING*. New York: Wiley.

Kutner, M. H., Nachtsheim, C. J., Neter, J., & Li, W. (2004). *Applied Linear Statistical Models*, (5ht ed.). IRWIN.

Lohr, S. L. (1999). *Sampling Design and Analysis*, (1st ed.). ITP.

Mafra, F. (2009). A avaliação do enquadramento em programas sociais a partir de características domiciliares.

Pessoa, D. G. C. & Silva, P. L. N. (1998). *Análise de Dados Amostrais Complexos*. IBGE.

SAS (2012). *SAS Institute Inc.* Cary, NC: SAS Institute Inc. SAS On Line Doc Version 9.3, `v9doc.sas.com`.

Silva, P. L. N., Pessoa, D. G. C., & Lila, M. F. (2003). Análise estatística de dados da pnad: Incorporando a estrutura do plano amostral. Technical report, Instituto Brasileiro de Geografia e Estatística - IBGE.

Särnal, C.-E., Swensson, B., & Wretman, J. (2003). *Model Assisted Survey Sampling*, (1st ed.). Springer Series in Statistics.

Apêndice A

- Macro SAS

```
%macro pnadsurveyreg(data=,y=,x=,intercept=YES,fpc=YES);
proc sort data=&data;
 by v4602 UPA;
run;
proc iml;
 use &data;
 read all var{&y} into y;
 read all var{&x} into x;
 read all var{v4729} into w;
 read all var{v4602} into h;
 read all var{UPA} into i;
 read all var{v4605} into fh1;
 read all var{v4607} into fh2;
 nomes={&x}‘;
 tab=y||x||w;
 ni=ncol(unique(i));
 nh=ncol(unique(h));
 n=nrow(y);
 sw=w[+];
 my=sum(y#w)/sw;
 mx=sum(x#w)/sw;
 %if %upcase(&intercept)=YES %then %do;
  x=j(n,1,1)||x;
  nomes={Intercept &x}‘;
 %end;
 beta=inv(x‘*(x#w))*x‘*(y#w);
 e=(y-x*beta)#w;
 e2=e#x;

/***** 1st stage *****/
 _h_=j(nrow(h),1,0);
 _h_[1]=1;
 _i_=j(nrow(i),1,0);
 _i_[1]=1;
 do k=2 to nrow(h);
  if h[k]=h[k-1] then _h_[k]=_h_[k-1];
  else _h_[k]=_h_[k-1]+1;
  if i[k]=i[k-1] then _i_[k]=_i_[k-1];
  else _i_[k]=_i_[k-1]+1;
 end;
tab1=tab||_h_||_i_;
 ehi=j(1,ncol(x)+4,1);
 ehi[1,3:2+ncol(x)]=e2[1,];
 ehi[1]=tab1[1,ncol(tab1)-1];
 ehi[2]=tab1[1,ncol(tab1)];
 ehi[ncol(ehi)-1]=
 fh1[1];
 do j=2 to nrow(tab1);
  if tab1[j,ncol(tab1)-1]=
  tab1[j-1,ncol(tab1)-1] & tab1[j,ncol(tab1)]=
  tab1[j-1,ncol(tab1)] then do;
  ehi[nrow(ehi),3:2+ncol(x)]=
  ehi[nrow(ehi),3:2+ncol(x)]+e2[j,];
  ehi[nrow(ehi),ncol(ehi)]=
  ehi[nrow(ehi),ncol(ehi)]+1;
  ehi[nrow(ehi),ncol(ehi)-1]=
  (ehi[nrow(ehi),ncol(ehi)-1]+fh1[j])/2;
 end;
  else ehi=ehi//(tab1[j,ncol(tab1)-1]||tab1[j,
  ncol(tab1)]||e2[j,]||fh1[j]||j(1,1,1));
 end;
 ee=
 ehi[,ncol(ehi)-1:ncol(ehi)];
 ehi=
 ehi[,1:ncol(ehi)-2]||j(nrow(ehi),3+ncol(x),0);
 ehi[1,3+ncol(x):2+2*ncol(x)]=ehi[1,3:2+ncol(x)];
 ehi[,ncol(ehi)-2:ncol(ehi)-1]=ee;
 ehi[1,ncol(ehi)]=ehi[1,ncol(ehi)-1];
 count=0;
 do j=2 to nrow(ehi);
  if ehi[j,1]=ehi[j-1,1] then do;
   ehi[j,3+ncol(x):2+2*ncol(x)]=
   ehi[j-1,3+ncol(x):2+2*ncol(x)]+ehi[j,3:2+ncol(x)];
   count=count+1;ehi[j,ncol(ehi)]=
   ehi[j-1,ncol(ehi)]+ehi[j,ncol(ehi)-1];
   ehi[j,ncol(ehi)-2]=ehi[j-1,ncol(ehi)-2];
  end;
  else do;
   if ehi[j-1,1]=1 then do;
   ehi[1:count+1,ncol(ehi)-1]=count+1;
   ehi[1:count+1,3+ncol(x):2+2*ncol(x)]=
   repeat(ehi[j-1,3+ncol(x):2+2*ncol(x)]/ehi[j-1,ncol(ehi)-1],count+1);
   in=ehi[j,2];
   ehi[1:count+1,ncol(ehi)]=
   repeat(ehi[j-1,ncol(ehi)],count+1);
  end;
   else do;
    ehi[in:in+count,ncol(ehi)-1]=count+1;
    ehi[in:in+count,3+ncol(x):2+2*ncol(x)]=
    repeat(ehi[j-1,3+ncol(x):2+2*ncol(x)]/ehi[j-1,ncol(ehi)-1],count+1);
    ehi[in:in+count,ncol(ehi)]=
    repeat(ehi[j-1,ncol(ehi)],count+1);
    in=ehi[j,2];
   end;
   ehi[j,3+ncol(x):2+2*ncol(x)]=
   ehi[j,3:2+ncol(x)];
   ehi[j,ncol(ehi)]=
   ehi[j,ncol(ehi)-1];
   count=0;
  end;
  if j=nrow(ehi) then do;
   ehi[in:in+count,ncol(ehi)-1]=count+1;
   ehi[in:in+count,3+ncol(x):2+2*ncol(x)]=
   repeat(ehi[j,3+ncol(x):2+2*ncol(x)]/ehi[j,
   ncol(ehi)-1],count+1);
   ehi[in:in+count,ncol(ehi)]=
   repeat(ehi[j,ncol(ehi)],count+1);
   in=ehi[j,2];
  end;
 end;
print 'Design Summary',,'Number of Strata' (ehi[nrow(ehi),1]),
'Number of Clusters' (ehi[nrow(ehi),2]);
*print ehi;
do jj=1 to nrow(ehi);
 if ehi[jj,ncol(ehi)-1]=1 then do;
```

```
  ehi[jj,ncol(ehi)-1]=2;
  ehi[jj,3:2*ncol(x)]=0;
  ehi[jj,3+ncol(x):2+2*ncol(x)]=0;
 end;
end;
G1=((n-1)/(n-ncol(x)))*(((ehi[,ncol(ehi)-1]/(ehi[,
ncol(ehi)-1]-1))#(ehi[,
3:2+ncol(x)]-ehi[,3+ncol(x):2+2*ncol(x)]))‘*(ehi[,
3:2+ncol(x)]-ehi[,3+ncol(x):2+2*ncol(x)]));
%if %upcase(&fpc)=YES %then %do;
G1=
((n-1)/(n-ncol(x)))*((((ehi[,ncol(ehi)-1]#(1-ehi[,
 ncol(ehi)-2]))/(ehi[,
 ncol(ehi)-1]-1))#(ehi[,3:2*ncol(x)]-ehi[,
 3+ncol(x):2+2*ncol(x)]))‘*(ehi[,
 3:2*ncol(x)]-ehi[,3+ncol(x):2+2*ncol(x)]));
%end;
varx1=vecdiag(inv(x‘*(x#w))*G1*inv(x‘*(x#w)));
gl=ehi[nrow(ehi),2]-ehi[nrow(ehi),1];

/***** 2nd stage *****/
 _h_=j(nrow(i),1,0);
 _h_[1]=1;
 do k=2 to nrow(i);
  if i[k]=i[k-1] then _h_[k]=_h_[k-1];
  else _h_[k]=_h_[k-1]+1;
 end;
 tab2=tab||_h_;
 ehi=j(1,2*ncol(x)+2,0);
 ehi[1,2:1+ncol(x)]=e2[1,];
 ehi[1]=tab2[1,ncol(tab2)];ehi[ncol(ehi)]=1;
 do j=2 to nrow(tab2);
  if tab2[j,ncol(tab2)]=tab2[j-1,ncol(tab2)] then do;
   ehi[nrow(ehi),2:1+ncol(x)]=ehi[nrow(ehi),2:1+ncol(x)]+e2[j,];
   ehi[nrow(ehi),ncol(ehi)]=ehi[nrow(ehi),ncol(ehi)]+1;
  end;
  else do;
   ehi=ehi//(tab2[j,ncol(tab2)]||e2[j,]||j(1,ncol(x)+1,1));
  end;
 end;
 ehi[,2+ncol(x):1+2*ncol(x)]=ehi[,2:1+ncol(x)]/ehi[,ncol(ehi)];
 do jj=1 to nrow(ehi);
  ehi2=ehi2//repeat(ehi[jj,2+ncol(x):2+2*ncol(x)],ehi[jj,ncol(ehi)]);
 end;
 eh=e2||ehi2;
%if %upcase(&fpc)=YES %then %do;
 eh=e2||ehi2[,1:ncol(ehi2)-1]||1-fh1#(1-fh2)||ehi2[,ncol(ehi2)];
%end;
do jj=1 to nrow(ehi);
 if ehi[jj,ncol(ehi)]=1 then ehi[jj,ncol(ehi)]=2;
end;
G2=
((n-1)/(n-ncol(x)))*(((eh[,ncol(eh)]/(eh[,ncol(eh)]-1))#(eh[,
1:ncol(x)]-eh[,1+ncol(x):2*ncol(x)]))‘*(eh[,
1:ncol(x)]-eh[,1+ncol(x):2*ncol(x)]));
%if %upcase(&fpc)=YES %then %do;
G2=((n-1)/(n-ncol(x)))*((((eh[,ncol(eh)]#(1-eh[,
ncol(eh)-1]))/(eh[,ncol(eh)]-1))#(eh[,1:ncol(x)]-eh[,
1+ncol(x):2*ncol(x)]))‘*(eh[,1:ncol(x)]-eh[,1+ncol(x):2*ncol(x)]));
%end;
varx2=vecdiag(inv(x‘*(x#w))*G2*inv(x‘*(x#w)));
stdx=sqrt(varx1+varx2);
t=beta/stdx;
probt=2*(1-probt(t,gl));
param=beta||stdx;
mse=sqrt((n*((y-x*beta)#w)‘*(y-x*beta))/((n-ncol(x))*sw));
print
"Number of Observations Used" n[label=’’],
"Sum of Weights" sw[label=’’],
"Weighted Mean of &y" my[label=’’],
"Root MSE" mse[label=’’];

print "Regression Analysis for Dependent Variable &y";
print "Estimated Regression Coefficients",,;
est={’Estimate’ ’Std. Error’};
print nomes[label=’Parameter’]param[format=comma12.8 colname=
est label=’’]t[format=comma6.2 colname=
’t Value’    label=’’]probt[format=pvalue6. colname=’P > |t|’ label=’’];
print
"NOTE: The denominator degrees of freedom for the t tests is" gl[label=’’]’.’;
quit;
%mend pnadsurveyreg;
%pnadsurveyreg(data=pes2007_1,y=remuneracao,x=anos_estudo);
%pnadsurveyreg(data=pes2007_1,y=remuneracao,x=anos_estudo,fpc=no);
%pnadsurveyreg(data=pes2007_1,y=remuneracao,x=anos_estudo,intercept=no);
%pnadsurveyreg(data=pes2007_1,y=remuneracao,x=anos_estudo,intercept=no,fpc=no);
```